Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- Season: Highest bike rentals happened during Fall, followed by Summer, Spring and Winter in order.
- Month: Demands were high for Jan, May, Jul, Aug & Oct, however when compared across years, Jun, Jul and Aug had the highest demands.
- Year: The demand increases with each passing year and retains the same pattern in ratio.
- workingday, holiday and weekday: they are talking about the same data however there is no clear pattern on which exact weekday is preferred nor is there any distinction on whether it maybe higher during a holiday or not.
- weathersit: Demand is there when the weather is clear and slightly more during light snow and rains during Fall and Winter seasons. All weather conditions during spring had the least demands.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

It is important to use drop_first=True during dummy variable creation to avoid multicollinearity and redundancy among them. This code will drop the first column from the dummy variable table and the data from the first column can be explained by other variables.

For example, assume that a categorical variable has 3 values namely A, B and C and using one-hot encoding this variable is converted to dummies and hence 3 columns are created. Now note the following:

- A can be represented as 100
- B can be represented as 010
- C can be represented as 001

However, this is redundant as A can be represented simply as 00 with B and C being represented as 10 and 01 respectively.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

temp has the highest correlation with the target variable with a value of 0.63.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Validation of the assumption of linear regression of the final model built on the training set was on the following:

- Linear relationship between dependent and independent variables using EDA and plotting of the actual and predicted values on the final model.
- Normal distribution of the residuals(errors) using histogram with mean hovering around 0.
- Independence of residuals from each other by plotting a scatter plot which shows no relationship between them.
- Homoscedasticity ie residuals have constant variance by plotting actuals and predicted values to confirm that there is no funnel pattern.
- Multicollinearity ie none of the variables have high multicollinearity between them by checking VIF.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top 3 features contributing significantly towards explaining the demand of the shared bikes are: year, temp and light snow/rain.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is used in machine learning to best linear-fit relationship on a given dataset between the dependent and independent variables and is based on supervised learning. It mostly uses sum of squared residuals method.

There are 2 types of linear regression.

- Simple linear regression: this is a relationship between one dependent and one independent variable using a straight line. The formula is as follows $Y=\beta_0+\beta_1X_1+\epsilon$ where Y is the dependent variable, β_0 is the constant, β_1 is the slope/coefficient, X_1 is the independent variable and ϵ is the error.
- Multiple linear regression: this is a relationship between one dependent and multiple independent variables being fit on to a hyperplane instead of a straight line. Its formula is $Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\epsilon$.

The equation for this best fit regression line can be found using 2 methods called Differentiation and Gradient descent.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is used to emphasise the important of visualization in data analytics over relying solely on statistics which can paint a completely different picture. Visualization shows a different dimension to the story of the data when datasets share similar statistics. It also warns of the dangers of outliers in datasets. Following is the proof of Anscombe's quarter.

It contains 4 datasets with each containing 11 pairs and each share the same descriptive statistics. See below.

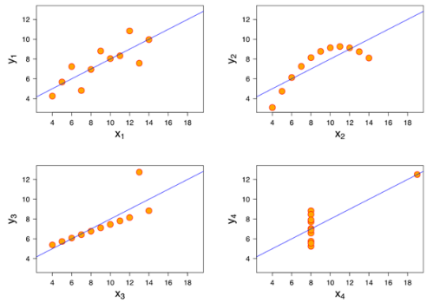
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Features of the above datasets are:

- Each dataset has a mean with x being 9 and mean of y being 7.50.
- Each dataset has a variance of x is 11 and y is 4.13.
- Each dataset has a correlation coefficient between x and y as 0.816 which is a strong positive correlation.

With this data alone, one may incorrectly conclude that all 4 datasets have the same behaviour and thus make incorrect inferences.

However, let us know plot these datasets to understand visually how these datasets actually appear.



Features to note are:

- Regression line is the same but the graphically representation of each datapoint is completely different.
- Dataset I looks to have a clean and well-fitted linear models.
- Dataset II is not distributed normally.
- Dataset III has a linear distribution but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as Pearson's correlation coefficient, is the measure of the strength of a linear association between two variables. It has range of -1 and 1. It's formula is as follows.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

It has 3 possible outcomes:

- Positive correlation: value greater than 0 and if 1 then it is total positive linear correlation. For example, relation between age and height.
- No correlation: value of 0. For example, relation between height and total marks scored in an exam.
- Negative correlation: value lesser than 0 and if it -1 then it is a total negative linear correlation. For example, relation between age and health.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method to bring the data of various variables within a particular range to execute further operations collectively.

Data range of features can vary between very low to very high values which effects the way machine learning weighs such values by weighing heavily on higher values thereby creating an inconsistent model that would determine incorrect coefficients. Scaling is used to ensure that these values are weighed correctly thereby building a stronger model.

Normalization is used to scale values in the range of [0,1]. The formula is given below.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardized scaling is used to scale values to have a mean 0 with a standard deviation of unit 1 variance. The formula is given below.

$$x = \frac{x - mean(x)}{sd(x)}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

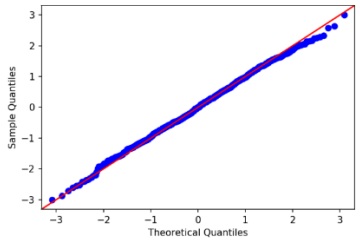
$$VIF_i = \frac{1}{1 - {R_i}^2}$$

Based on VIF formula, VIF will be infinite only when R2 is 1 and this happens when there is perfect correlation between all the features/variables in the current model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q plot, also known as Quantile-Quantile plot, is a graphical plot created by plotting two sets of quantiles against one.

It is used for determining if two datasets come from populations with a common distribution, whether they have a common location and scale, whether they have similar distributional shapes or similar tail behaviour.



The purpose of this plot can be seen in the graph above of an ideal Q-Q plot, the quantiles are samples from the same distribution and therefore roughly fall on a 45-degree straight line. However, if the quantile samples were not from the same distribution, then there would be a greater departure from this reference line.